Application of Discrete Discriminant Analysis In Public Health Phenomenon A Safe-Motherhood Study

Kuntoro Kuntoro[1], Rahmah Indawati[1], Diah Indriani[1], Arief Wibowo[1]

[1]Department of Biostatistics and Population Study, Airlangga University School of Public Health, Surabaya 60115, Indonesia.
kuntoro1@indo.net.id, kuntoro2@yahoo.com; 1; 1; 1

Abstract

Linear discriminant analysis is used widely in medical sciences since most discriminator variables have interval and ratio scale besides multivariate normal and unknown but equal dispersion and covariance matrices assumptions are met. However, in public health sciences most discriminator variables have ordinal and nominal scales so that linear discriminant analysis is not recommended. When ordinal and nominal scales are categorized into positive and negative, a discrete discriminant analysis is prefered.

The data of safe-motherhood collected by Kuntoro et al. (1999) that include group assignment (mother : dead, alive), five discriminator variables $(\mathbf{x})$, place of delivery baby, person who delivered baby, antenatal care, place of antenatal care, and referal services obtained. Allocation of a new observation x into Group 1 (dead) if $\hat{g}_1(\mathbf{x}) > \hat{g}_2(\mathbf{x})$ and into Group 2 (alive) if $\hat{g}_1(\mathbf{x}) < \hat{g}_2(\mathbf{x})$.

The results showed that in the beginning of group assignment there were 65 mothers who were dead and 976 mothers who were alive. Discrete discriminant analysis gives 93.95% correct classification, while linear discriminant analysis, logistic regression analysis, and neural network discriminant analysis give respectively 86.6%, 93.76% and 93.76%.

It is concluded that discrete discriminant analysis seems to be the most effective in discriminating individuals into groups when the discriminator variables are dichotomously categorical.

Keywords: correct classification, discriminator, group assignment

1 Introduction

Determining the procedure for classifying an individual who has a number of physical and laboratory examination results falls into a disease or non disease under study is important in medical sciences. This procedure is required by a medical professional for selecting the medical action when an individual falls into a disease group under study. This procedure is commonly called **cut-off**.

There are several statistical methods related to **cut-off**, one of them is linear discriminant analysis. It is used widely in medical sciences since most discriminator variables have interval and ratio scale besides multivariate normal and unknown but equal dispersion and covariance matrices assumptions are met.

In public health sciences as well as in medical sciences the procedure mentioned above is also required by a public health professional for the same purpose. Unlike in medical sciences whose variables of interest have interval or ratio scales, in public health sciences whose variables of interest have ordinal and nominal scales. In this case, a linear discriminant analysis is not recommended. When ordinal and nominal scales are categorized into positive and negative, a discrete discriminant analysis is prefered.

Four statistical methods are discussed, a discrete discriminant analysis, a linear discriminant analysis, logistic regression analysis, and neural network analysis. The degrees of correct classification are compared.

2 Basic Concepts

2.1 Discrete Discriminant Analysis

Consider random variables $X_1, X_2, \ldots, X_p$. It is assumed that each random variable has at most a finite number of distinct values $s_1, s_2, \ldots, s_p$. Moreover, the product represents the set of all possible values. Furthermore, x denotes a particular realization of X. The underlying distribution within each subpopulation is assumed to be multinomial. In the case of two groups, let the class conditional density in G_i be f_i and the prior probability of membership be P_i where $i = 1, 2$. A new observation x will be assigned into G_1 if $P_1 f_1(x) > P_2 f_2(x)$ and into G_2 if $P_1 f_1(x) < P_2 f_2(x)$. The assignment is done randomly if $P_1 f_1(x) = P_2 f_2(x)$ [2] Let $P_1 f_1(x)$ be a discriminant score, its value is denoted by $g_1(x)$. Let n observations be samples from the combined population, and let $n_i(x)$ be the number from G_i having measurment x. Hence, $n = \sum_x \sum_i n_i(x)$. Let $P_i \frac{n_i}{n}$ and $\hat{f}_i(x) = \frac{n_i x}{n n_i}$ be respectively nonparametric estimates of prior probabilities and the class conditional densities that yield sample-based estimates of the discriminant scores, [2] The sample-based rule can be described as follows : Assign x to G_1 if $\hat{g_1}(x) > \hat{g_2}(x)$ and to G_2 if $\hat{g_1}(x) < \hat{g_2}(x)$ and allocate randomly if $\hat{g_1}(x) = \hat{g_2}(x)$ [2]

2.2 Linear Discriminant Analysis

Let $D_i = b_1 x_1 i + b_2 x_2 i + ... + b_p x_p i$ be standardized linear discriminant function, where D_i is score discriminant for individual i, b_i is discriminant coefficient, where $i = 1, 2, \ldots, p$ and x_i is discriminator variable, where $i = 1, 2, \ldots, p$. Let $\frac{\bar{D_1} + \bar{D_2}}{2}$ be a cut-off, where $\bar{D_1}$ and $\bar{D_2}$ are the means of discriminant scores in respectively group 1 and group 2. The rule of classification: If $\bar{D_1} > \bar{D_2}$, an individual with a given value of x_i is assigned to G_1 if his/her discriminant score $D_i >$ cut-off and to G_2 if his/her discriminant score $D_i <$ cut-off. If $\bar{D_1} < \bar{D_2}$, an individual with a given value of x_i is assigned to G_1 if his/her discriminant score $<$ cut-off and to G_2 if his/her discriminant score $>$ cut-off [3, 4]

2.3 Logistic Regression Analysis

Logistic Regression Model relates a number of independent variables whose scales can be nominal, ordinal, interval and ratio and a dependent variabel whose scale is nominal categorized into two classes, yes or no called binary. Logistic regression model has similarity to linear discrimant model for two groups when its independent variables have interval and/or ratio scale, while its dependent variable has nominal scale with two categories. The formula can described as follows.

$E(Y) = \frac{1}{1+exp[-(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)]}$

Where $E(Y)$ is the expected value of Y, β_0 and β_1, $p = 1, 2, \ldots$ are logistic regression coefficients. Moreover, $E(Y)$ is equivalent to the probability $Pr(Y = 1)$ for $(0, 1)$ random variables such as Y. Hence, the probability that one of the two possible outcomes of Y occurs can be described as follows [3, 4].

$Pr(Y = 1) = \frac{1}{1+exp[-(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)]}$

For example, given the values of independent variables such as educational level of mother, adequacy food consumption to child, food expenditure for child, we can compute the probability that a child falls in malnutrition condition $(Y = 1)$ or he/she falls in normal condition $(Y = 0)$.

2.4 Neural Network Analysis

The following formula shows a discrimination function as a linear combination of x variables.

$g(\mathbf{x} = \mathbf{w}^T + w_0$

where $\mathbf{x}$ is a vector of discriminator variables, $\mathbf{w}^T$ is the transpose weight vector and w_0 the bias or threshold weight. In a case of two categories, it is used the following decision rule : If $g(\mathbf{x}) > 0$ then decide w_1 If $g(\mathbf{x}) < 0$ then decide w_2 That means if the inner product $\mathbf{w}^T\mathbf{x}$ exceeds the threshold w_0 , $\mathbf{x}$ is assigned to w_1 and to w_2 otherwise. Moreover, $\mathbf{x}$ can ordinarily be assigned to either class, or can be left undefined if $g(\mathbf{x}) = 0$. Furthermore,

Table 1: Table 1 Procedure for Computing Class Condition Densities

	G_1 (DEAD)		G_2 (ALIVE)	
CELL x_1, x_2, x_3, x_4, x_5	O_1	$\hat{f}_1$	O_2	$\hat{f}_2$
1,1,1,1,1	7	0.108	331	0.331
1,1,1,0,1	1		6	
1,1,1,0,0	0		1	
0,1,0,1,1	0		2	
Total	65		976	

the decision surface is defined by the equation $g(\mathbf{x}) = 0$. The decision surface separates points assigned to w_1 from points assigned to w_2. The decision surface is a hyperplane when $g(\mathbf{x})$ is linear.

If $\mathbf{x}_1$ and $\mathbf{x}_2$ are both on the decision surface, then $\mathbf{w}^T\mathbf{x}_1 + w_0 = \mathbf{w}^T\mathbf{x}_2 + w_0$ or $\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0$ [1]

$\mathbf{w}$ is normal to any vector that lies in the hyperplane. The feature space is divided by the hyperplane into two half-spaces: decision region R_1 for w_1 and region R_2 for w_2. Since $g(\mathbf{x}) > 0$ if $\mathbf{x}$ is in R_1, it follows that the normal vector $\mathbf{w}$ points into R_1.

3 Material and Methods

3.1 Material

To show the computation of discrete discriminant model, the secondary data of safe-motherhood collected by [5] that include group assignment (mother : dead, alive), five discriminator variables (x) such as place of delivery baby (x_1), person who delivered baby (x_2), antenatal care (x_3), place of antenatal care (x_4), and referal services obtained (x_5). The total number of subjects to be analyzed is 1041 persons, among them 65 persons were dead and 976 persons were alive.

3.2 Methods

The class condition densities in G_1 and G_2 are computed by formula as follows: $\hat{f}_i(x) = \frac{n_i(x)}{n_i}$

For example, $\hat{f}_1(1,1,1,1,1) = \frac{n_1(1,1,1,1,1)}{n_1} = \frac{7}{65} = 0.100$ $\hat{f}_2(1,1,1,1,1) = \frac{n_2(1,1,1,1,1)}{n_2} = \frac{331}{976} = 0.339$

The discrimination score in G_1 and G_2 are computed by formula as follows: $\hat{g}_i(x) = \frac{n_i}{n}\frac{n_i(x)}{n_i} = \frac{n_i(x)}{n}$

For example, discriminant score of individual with $x(1,1,1,1,1)$ in group 1 is $\hat{g}_1(1,1,1,1,1) = \frac{n_1(1,1,1,1,1)}{n} = \frac{7}{1041} = 0.0007$

For example, discriminant score of individual with $x(1,1,1,1,1)$ in group 2 is $\hat{g}_2(1,1,1,1,1) = \frac{n_2(1,1,1,1,1)}{n} = \frac{331}{1041} = 0.318$

Allocation of a new observation $\mathbf{x}$ into Group 1 (dead) if $\hat{g}_1(\mathbf{x}) > \hat{g}_2(\mathbf{x})$ and into Group 2 (alive) if $\hat{g}_1(\mathbf{x}) < \hat{g}_2(\mathbf{x})$. In the example mentioned above, since $\hat{g}_1(1,1,1,1,1) = 0.0007 < \hat{g}(1,1,1,1,1) = 0.318$, then the new observation falls into Group 2 or alive.

For demonstrating three other statistical analysis, the data were analyzed by means of statistical soft-ware such as SPSS and S-Plus.

4 (Results and Discussion

4.1 Discrete Discriminant Analysis

Among 65 individuals who were the members of Group 1, 63 individuals were misclassified, 2 individuals were correctly classified. Among 976 individuals who were the members of Group 2, none was misclassified, then all of them were correctly classified.

Table 2: Table 2 Discriminant Scores and Misclassification

CELL x_1, x_2, x_3, x_4, x_5	G_1 O_1	G_1 $\hat{f}_1$	G_2 O_2	G_2 $\hat{f}_2$	Discriminant Score $\hat{g}_1$	Discriminant Score misclassification	Discriminant Score $\hat{g}_2$	Discriminant Score misclassification
1,1,1,1,1	7	0.108	331	0.339	0.007	+	0.318	-
1,1,1,0,1	1	0.015	6	0.006	0.001	+	0.006	-
1,1,1,0,0	0	0.000	1	0.001	0.000	+	0.001	-
1,1,0,1,1	1	0.015	179	0.183	0.001	+	0.172	-
1,1,0,1,0	1	0.015	1	0.001	0.001	+	0.001	-
1,1,0,0,1	26	0.400	356	0.365	0.025	+	0.342	-
1,1,0,0,0	10	0.154	44	0.045	0.010	+	0.042	-
1,0,1,1,1	5	0.077	26	0.027	0.005	+	0.025	-
1,0,1,0,1	0	0.000	1	0.001	0.000	+	0.001	-
0,1,1,1,1	0	0.000	2	0.002	0.000	+	0.002	-
0,1,1,1,0	2	0.031	0	0.000	0.002	-	0.000	+
1,0,0,1,1	3	0.046	4	0.004	0.003	+	0.004	-
1,0,0,0,1	4	0.062	7	0.007	0.004	+	0.007	-
0,1,0,0,1	2	0.031	7	0.007	0.002	+	0.007	-
0,1,0,0,0	3	0.046	9	0.009	0.003	+	0.009	-
0,1,0,1,1	0	0.000	2	0.002	0.000	+	0.002	-
Total	65		976					

Table 3: Table 3 Degree of Misclassification Using Linear Discriminant Analysis

Original Group Membership	Prediction Group Membership G_1 (Dead)	Prediction Group Membership G_2 (Alive)	Total
G_1 (Dead)	30	35	65
G_2 (Alive)	104	872	976
Total	134	907	1041

The degree of misclassification is (63+0)/1041 = 6.05
Hence, the degree of correct classification is $\frac{(2+976)}{1041} = 93.95\%$

4.2 Linear Discriminant Analysis

The degree of misclassification is $\frac{(35+104)}{1041} = 13.4\%$ The degree of correct classification is 86.6

4.3 Logistic Regression Analysis

The degree of misclassification is $\frac{(65)}{1041} = 6.24\%$ The degree of correct classification is 93.76

4.4 Neural Network Analysis

The degree of misclassification is $\frac{(65)}{1041} = 6.24\%$ The degree of correct classification is 93.76

Table 4: Table 4 Degree of Misclassification Using Logistic Regression Analysis

Observed Mother Status	Prediction Mother Status G_1 (Dead)	Prediction Mother Status G_2 (Alive)	Total
G_1 (Dead)	0	65	65
G_2 (Alive)	0	976	976
Total	0	1041	1041

Table 5: Table 5 Degree of Misclassification Using Neural Network Analysis

Observed Mother Status	Prediction Mother Status		Total
	G_1 (Dead)	G_2 (Alive)	
G_1 (Dead)	0	65	65
G_2 (Alive)	0	976	976
Total	0	1041	1041

The first method gives the highest degree of correct classification. However, it is time consumed to compute for researchers who are not familiary with mathematical computation. For those who understand the program language such as Fortran, S-plus etc., they can develop statistical software so that it can reduce time operation as well as the well-known statistical software. The second method gives the lowest degree of correct classification. The third and fourth methods gives the same degree of correct classification. However, the fourth method requires the understanding about program language such as S- Plus. The third method is preferred by researchers who are familiary with statistical software such SPSS, SAS, MINITAB.

5 Conclusion and Recommendation

Discrete discriminant analysis gives 93.95% correct classification, while linear discriminant analysis, logistic regression analysis, and neural network discriminant analysis give respectively 86.6Discrete discriminant analysis seems to be the most effective in discriminating individuals into groups when the discriminator variables are dichotomously categorical.

Discrete discriminant analysis is recommended for researchers who feel convenient in doing mathematical calculation using spreadsheet software. However, for those who prefer easy statistical software, logistic regression is recommended.

References

[1] Ömer Cengiz ÇELEBÍ *http://www.celebisoftware.com/Tutorials/neural_networks/ accessed in 30 June 2009 at 7.50pm.* Celebisoftware (2009)

[2] W.R. Dillon and M. Goldstein, *Multivariate Analysis Methods and Applications*, John Wiley and Sons, Inc., New York (1984).

[3] D.G. Kleinbaum , L.L. Kupper and K.E. Muller, *Applied Regression Analysis and Other Multivariable Methods*, PWS-Kent Publishing Company, Boston (1988).

[4] D.G. Kleinbaum , L.L. Kupper, K.E. Muller and A.Nizam *Applied Regression Analysis and Other Multivariable Methods*, Duxbury Press, Pacific Grove (1998).

[5] K. Kuntoro, A. Wibowo, R. Indawati, M. Mahmudah, S. Melania, M.B. Qomaruddin, L. Sulistyorini *Longitudinal Study of Implementation of Social Security Net - Health Sector in East Java Province Indonesia*, Airlangga University, Surabaya (1999)